Structural genomics: beyond the Human Genome Project

Stephen K. Burley[1,2], Steven C. Almo[3], Jeffrey B. Bonanno[1,2], Malcolm Capel[4], Mark R. Chance[3], Terry Gaasterland[2], Dawei Lin[4], Andrej Šali[2], F. William Studier[4] & Subramanyam Swaminathan[4]

With access to whole genome sequences for various organisms and imminent completion of the Human Genome Project, the entire process of discovery in molecular and cellular biology is poised to change. Massively parallel measurement strategies promise to revolutionize how we study and ultimately understand the complex biochemical circuitry responsible for controlling normal development, physiologic homeostasis and disease processes. This information explosion is also providing the foundation for an important new initiative in structural biology. We are about to embark on a program of high-throughput X-ray crystallography aimed at developing a comprehensive mechanistic understanding of normal and abnormal human and microbial physiology at the molecular level. We present the rationale for creation of a structural genomics initiative, recount the efforts of ongoing structural genomics pilot studies, and detail the lofty goals, technical challenges and pitfalls facing structural biologists.

Rationale for the structural genomics initiative

Since Linus Pauling described sickle cell anaemia as the first "molecular disease", there has been an explosive growth in our understanding of the precise molecular mechanisms responsible for a significant number of human disease processes. In large part, these advances can be attributed to generous public and private support of investigator-initiated, hypothesis-driven research in North America, Europe and Asia. Biomedical researchers throughout the world are now busy establishing a new paradigm for human disease, one that implicates individual biological macromolecules. To paraphrase Pasteur, Koch and other great microbiologists of the late 19th and early 20th centuries, modern-day molecular and cellular biologists have become proponents of a "gene product theory of human disease." Instead of examining microbial invaders, the biomedical research community is studying the consequences of introducing foreign proteins (bacterial, fungal and viral virulence factors) into humans, the results of individual genetic lesions (gain/loss of function, alterations in function), or the cumulative effects of multiple genetic factors contributing to diseases such as adult-onset diabetes mellitus, hypertension and so on. This molecular view of disease has also contributed to the newly realized importance of three-dimensional structural studies by X-ray crystallography and solution nuclear magnetic resonance (NMR) spectroscopy. In favourable cases, it has even been possible to determine the structures of wild-type and mutant proteins implicated in human disease.

A strategic investment made in high-throughput genome sequencing, a 'big' science endeavour relatively new to biology, is also contributing to dramatic changes in our thinking. Using software packages, such as MAGPIE (http://genomes.rockefeller.edu/magpie/magpie.html/), we can compare organisms at the level of whole genomes, gleaning important evolutionary insights and identifying clinically relevant differences between man and viral/bacterial/fungal pathogens. The availability of whole-genome sequences also creates the potential to develop massively parallel tools, such as arrays of immobilized DNA elements to study gene expression patterns, that will contribute to both fundamental research and point-of-care diagnostics. Finally, newly characterized gene products themselves offer the promise of novel therapeutic agents, many of which will become protein pharmaceuticals.

As the Human Genome Project revved up, accelerating the pace of discovery in biology, technical advances increased the speed with which we can determine the three-dimensional structures of biological macromolecules. PCR-based recombinant DNA technology, high-level protein expression systems, robotic crystallization, cryogenic crystal handling, X-ray area detectors, high-field NMR spectrometers, tunable synchrotron radiation sources and high-performance computing have together catapulted structural biology from an esoteric niche to the biological mainstream. Structure determinations that used to require large teams gutting out a 20 person-year effort now constitute a single chapter in a graduate student's doctoral thesis.

High-speed computing has also revolutionized what we can do with this wealth of structural information. The impact of computational biology can be attributed, at least in part, to the modest complexity of protein fold space, which stands in stark contrast to the Byzantine character of the 100,000 (or so) genes encoded by the human genome. Although there is still some uncertainty regarding precise numbers, we now appreciate that the universe of compact globular protein folds is quite limited. Current estimates suggest that there are between 1,000 and 5,000 distinct spa-

[1]Howard Hughes Medical Institute, [2]The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. [3]Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. [4]Department of Biology, Brookhaven National Laboratory, Upton, New York 11973, USA. Correspondence should be addressed to S.K.B. (e-mail: burley@rockvax.rockefeller.edu).

tial arrangements of polypeptide chains found in nature[1]. The Protein Data Bank (PDB, http://www.rcsb.org/pdb/) contains three-dimensional structures of only about 800 distinct protein folds, with some 'popular' folds such as the eightfold αβ barrel of the triose phosphate isomerase (TIM) type being represented by 19 protein superfamilies (Fig. 1). In eukaryotes, most genes encode proteins with multiple globular domains (the average domain size is 153 (±87) residues[2]), giving many larger proteins the appearance of beads on a string (Fig. 2). Typically, a single 'bead' is responsible for carrying out a specialized biochemical task, such as phosphorylation of protein substrates by the kinase domain of the Src oncoprotein, which also contains SH2 and SH3 domains that are respectively responsible for binding phosphotyrosyl and poly-proline peptides. A significant evolutionary change in gene sequence often manifests itself at the level of an individual protein functional unit or domain, which may be regarded as a focal point of natural selection. If the resulting amino acid substitution destabilizes the structure of a critical domain in an essential gene product, survival is compromised and the mutated gene will disappear. When, however, residues in the active site of a functional domain are changed to create a useful new biochemical activity, the genetically altered organism stands to benefit and the mutated gene may persist.

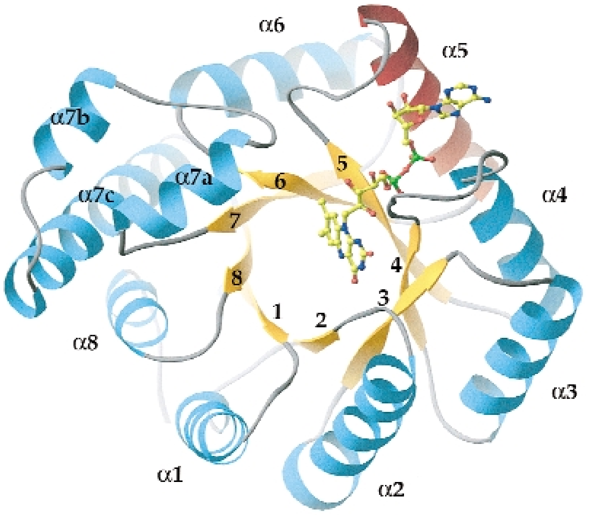

Fig. 1 Ribbon drawing of the structure of *E. coli* methylenetetrahydrofolate reductase, a single-domain protein with a T/M barrel that binds the cofactor flavin adenine dinucleotide (stick figure; reproduced from ref. 20).

It is somewhat *passé* to be advocating the potential benefits of combining three-dimensional structures with the results of the Human Genome Project. The limited number of protein structures deposited into the PDB has already proven extremely useful, taking us well beyond the evolutionary implications described in the previous paragraph. Protein fold assignment and homology modelling of related protein structures have become important research tools, providing structural insights for many different areas of biology and medicine. Large-scale protein structure analyses have even been applied to whole genomes, including *Mycoplasma genitalium*[3–5], *Saccharomyces cerevisiae*[6], *Escherichia coli* and *Caenorhabditis elegans* (A.Š., unpublished data). Although the results of these systematic efforts are encouraging to say the least, the paucity of the protein structure database limits the scope of modelling activities to 42% of the ORFs of the yeast genome (composed of over 6,000 genes). If one considers that only a fraction of a protein can usually be modelled, the situation looks decidedly worse (18% of all residues or domains in yeast proteins).

The obvious solution to this problem is to obtain complete three-dimensional structural information for each distinct protein fold. The best strategy for achieving this goal in a timely fashion needs to be rigorously justified. *De novo* prediction of a protein structure from its seqeunce is simply not feasible at present, forcing us to rely on experimentally determined structures. Eventual success of an experimental program is not in doubt (although it may take decades under the business-as-usual paradigm). Protein crystallographers and NMR spectroscopists are currently producing more than 6,000 structures per year. An exhaustive analysis[1] of all 1,107 new PDB submissions in 1994 revealed the following breakdown: 70% had essentially the same sequence as an existing PDB entry, 21% were closely related to an existing PDB entry and 9% had no obvious homologue in the PDB. Not surprisingly, only about one-third of the proteins constituting this 9% represented completely new folds. Thus, even at the current rate of 6,000 PDB structure depositions per year, structural biologists will probably not realize more than about 200 new protein folds annually. The spectacular success of the Human Genome Project suggests that a systematic effort in high-throughput structural genomics may be able to do it better, faster and cheaper. Unfortunately, the analogy with genomic sequencing is not without flaws.

As a practical matter, we cannot launch a structural genomics initiative aimed at determining the three-dimensional structure of every protein encoded by the human genome. Such an undertaking would take decades. It would also yield a considerable number of redundant structures, reflecting the relatively small size of protein fold space. Current estimates suggest that a directed program of structural study focusing on 10,000 selected domain targets would yield examples of virtually every distinct globular protein fold[6]. The result would be an appropriately redundant database of structures capable of supporting homology modelling of reasonable quality for nearly every globular segment of every protein found in nature. Within the structural genomics community itself, debate centres on the nature of this target list. Outside this small group, opinion runs the gamut from staunch opposition to high enthusiasm.

To promote further discussion of the technical and strategic problems posed by a structural genomics initiative, we consider six logistical questions: (i) Can it be done?; (ii) Which targets should be chosen?; (ii) How will it be done?; (iv) Who should do it and at what cost?; (v) When will the task be substantially complete?; and (vi) What collateral benefits will accrue to biomedicine? Structural Genomics Workshops conducted by the National Institute of General Medical Sciences (NIGMS) have also examined some of these issues, and their reports are available (http://www.nih.gov/nigms/funding/psi.html).

Can it be done?

The feasibility of a large-scale, high-throughput structure determination program modelled on the Human Genome Project is being explored by a number of pilot studies underway in North America, Europe and Asia. Efforts are being made using both X-ray crystallography and solution NMR, with target lists derived from the genome sequences of archaebacteria (http://www.doe-mbi.ucla.edu/ProteomicsConsortium, http://www.riken.go.jp/KENCHO/GENOME/indexE.html/), eubacteria (http://s2f.carb.nist.gov/) and eukaryotes (http://proteome.bnl.gov/, http://

www.nmr.cabm.rutgers.edu/structuralgenomics/). Most of these groups have already deposited structures to the PDB, and there is general agreement within the structural biology community that all of the necessary basic technologies are in place, at least for the X-ray crystallographic approach. Within the United States, communication among the various pilot projects has been streamlined with the creation of two federally funded web sites (http://structuralgenomics.org/, http://presage.stanford.edu/). Recently, the NIGMS issued a request for applications for special grants to fund expansion of a small number of structural genomics pilot studies (http://www.nih.gov/nigms/funding/psi.html), which should permit further exploration of the logistical and financial consequences of a structural genomics initiative.

Which targets should be chosen?

Target selection is the most important strategic issue confronting the groups pursuing structural genomics pilot studies. Their respective performances will be measured in terms of the number of structures determined, what fraction contain novel folds, their impact on biology, and the cost per structure. The question of which structures to target is also relevant to the US funding agencies, because the US taxpayer and Congress will need to be convinced of the social and medical benefits of a structural genomics initiative before the required funds can be made available. Numerous discussions of target selection have been held[7,8]; (http://lion.cabm.rutgers.edu/bioinformatics_meeting/, http://www.nih.gov/nigms/funding/psi.html) and there is general agreement that it represents a highly technical research problem in its own right. There is no such agreement, however, on the extent to which biomedical criteria should play a role in target selection. Most groups working in structural genomics are developing target lists using small bacterial genomes on the grounds that these candidates will be both technically feasible and likely to provide reasonable coverage of the universe of protein folds. Their goal is to fill in the database of protein folds as soon as possible.

The pilot study being carried out by the New York Structural Genomics Research Consortium (NY-SGRC, including Albert Einstein College of Medicine, Brookhaven National Laboratory, Mount Sinai School of Medicine, The Rockefeller University and Weill Medical College of Cornell University) is taking a different tack (http://proteome.bnl.gov/). We have taken the position that the target list should be designed to go beyond the problem of completing the database of protein folds. Wherever possible, we will select structure determination candidates that correspond to human proteins (and their yeast homologues) implicated in the causes and treatment of disease, plus fungal, bacterial and viral virulence factors. We believe that our pilot study will yield protein structures that are immediately useful in trying to understand and treat human disease. Our program will also produce substantial quantities of a large number of disease gene proteins for exhaustive biochemical characterization, high-throughput screening programs, and discovery and optimization of lead compounds.

How will it be done?

There is still considerable debate within the nascent structural genomics community regarding the best approach to the problem, and space considerations preclude making a detailed comparison. Instead, we describe the initial experimental design, progress to date and long-term plans of the NY-SGRC. Given our long-term commitment to focus on human disease gene proteins, we have selected a simple, well-studied eukaryote (*S. cerevisiae*) as our first target organism. Structures will be determined by X-ray crystallography using synchrotron radiation. Only this approach is suitable for high-throughput data collection (compare 30 minutes using a third generation synchrotron source to 30 days with an NMR spectrometer).

The goal of this phase of the NY-SGRC pilot study is to develop the technology to routinely carry through the entire process of obtaining protein structures, starting from their gene sequences. We seek an appreciation of the types of problems, the likely success rate and the feasibility of large-scale, highly parallel structure determination. The process (Fig. 3) involves: (i) PCR amplification of the coding sequence from genomic or cDNA; (ii) cloning the coding sequence into an appropriate expression vector; (iii) expressing the protein at a sufficiently high level; (iv) sequencing the cloned gene to verify that the coding sequence was correctly amplified; (v) confirming the identity of the expressed protein and characterizing it to establish the likelihood of crystallizability; (vi) obtaining the protein in sufficient amounts and purity to form crystals; (vii) defining crystallization conditions; (viii) labelling protein with selenomethionine, purifying it, and obtaining and freezing diffraction-quality crystals for X-ray crystallography by the technique of multiple-wavelength anomalous dispersion (MAD); (ix) collecting MAD data at an X-ray beamline (National Synchrotron Light Source, Brookhaven National Laboratory); (x) determining the phases of the reflections, building the model and refining the structure; and (xi) making functional inferences from the structure and disseminating our findings. Failures were anticipated at every step, making the process somewhat akin to a funnel, with a broad input and narrow output.

For simplicity, each target protein has a unique sequential identifier, P001–P018 for the targets that have already been cloned and expressed plus P019–P111 for the 93 proteins that are currently being processed, with 3 controls, in 96-well format to test procedures for scale-up. From the outset, we have posted our target selections and progress toward structure determination on our publicly accessible web site (http://proteome.bnl.gov/). The web page has links to the *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces/), SwissProt (http://expasy.hcuge.ch/), ProDom (http://www.toulouse.inra.fr/prodom.html/) and OMIM (http://www.ncbi.nlm.nih.gov/Omim) for information about the target proteins and their protein families, links to the output of a BLAST search of the yeast protein against the nonredundant database of protein sequences, and links to information about related human proteins (“Human Gene Information”).

For our initial test, we selected 12 yeast proteins for which no structural information was available or predicted (P001–P012). We also selected four domains for which structures had been predicted with high confidence[6], even though the amino acid identities with known structures were low (14–22%). These first 16 candidates were cloned and expressed as a group. Two more proteins, P017 and P018, were added later. Significant *E. coli* expression with T7 RNA polymerase[9] was achieved for all of the 18 targets. Of the first 18 proteins, only P010 and P011 were too insoluble for further work. Initial crystallization trials were attempted with P001–P009, P012, P017 and P018. Manual crystallization quickly became a serious bottleneck, which was overcome using crystallization robotics. Of 12 purified proteins tested, 11 gave microcrystals under one or more conditions in our initial screens.

Despite the remarkably high success rate for obtaining microcrystals, routine production of diffraction-quality crystals has proved to be more challenging. To date, we have obtained structures of P007 and P008 (PDB codes: 1b54, 1ci0). The results of homology modelling with each newly obtained structural template represent an important benefit of any structural determination. We performed automated protein structure modelling with both P007 and P008, respectively, producing 20 and 11 good-quality homology models of proteins for which structural information was not previously available. The resulting homology models can be found in MODBASE (http://guitar.rockefeller.edu/modbase/).

To our chagrin, P007 turned out to be an example of the "popular" TIM barrel fold described earlier. The fact that we could construct homology models for 20 new protein sequences with the structure of P007 is, nonetheless, reassuring, because it documents that structural genomics will be of some value even when determining an "unknown" structure does not add a new fold to the PDB. P007, a protein of unknown function, most closely resembles the amino-terminal domain of an alanine racemase (PDB code: 1sft). The similarity of the active sites of the two proteins suggests that P007 is an amino acid racemase, and preliminary biochemical assays indicate that P007 does in fact have such an activity.

P008 represents a novel protein fold, with PNP oxidase activity. In collaboration with John Blanchard of Albert Einstein College of Medicine, we are addressing a number of important questions pertaining to its precise enzymatic mechanism. The structure of P008 raises an interesting issue that will influence target selection for structural genomics. Twenty-five per cent of P008 resembles a flavin mononucleotide (FMN)-binding protein from *Desulfovibrio vulgaris*, which was subsequently determined by solution NMR (PDB code: 1axj). P008 and the FMN-binding protein cannot, however, be said to have the same fold, because P008 has many additional secondary structural elements. It may be more accurate to say that they are close to one another in the continuum of protein fold space. Sander has referred to the concept of "attractors" in fold space, and this may be a useful way to look at the problem[10]. The goal of trying to complete the elucidation of all distinct protein folds will be complicated by such examples, and the search for the best way of dealing with this situation represents a fertile area of research in bioinformatics.

P001–P006, P009 and P012 have been characterized using limited proteolysis combined with mass spectrometry to identify disordered regions that may be interfering with formation of large crystals[11–13]. So far, four of the proteins have demonstrated rapid proteolysis of N- and carboxy-terminal regions with protease-resistant core domains. Subclones encoding appropriately truncated forms of these target proteins have been prepared, and protein expression, purification and crystallization are underway.

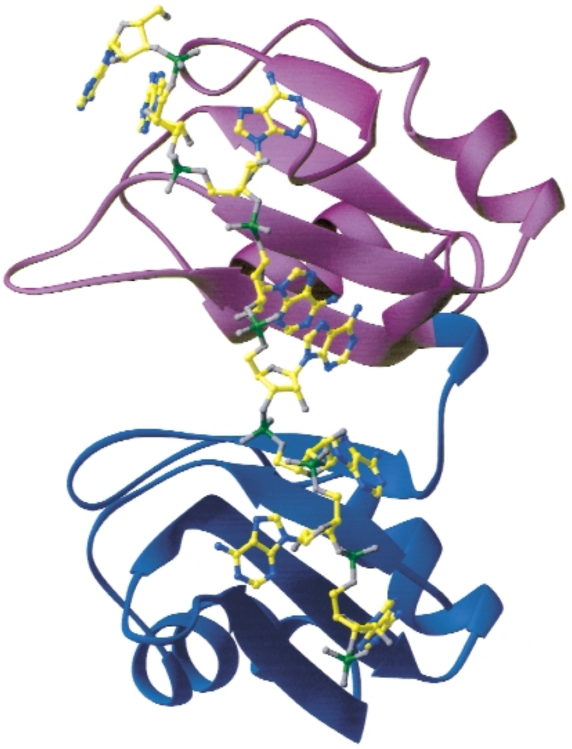

Fig. 2 Unlike the single-domain structure depicted in Fig. 1, the PABP consists of multiple domains (RNA-recognition motifs, depicted in blue and purple), which together create a continuous RNA-binding surface that interacts with polyadenylate RNA (stick figure; reproduced from ref. 21).

Our experience at the one-year mark with the first 18 targets from yeast documents that the techniques we are implementing will support a structural genomics initiative. The process resembles a funnel, albeit one with a surprisingly good throughput. The genes for all 18 targets (100%) were successfully cloned and expressed in *E. coli,* and only 2 were too insoluble to work with. Of the 12 reasonably soluble target proteins pursued further, all could be purified in significant quantities, and 11 of 12 yielded microcrystals. Structures have been determined for two proteins, and their atomic coordinates deposited in the PDB. Analyses of the other crystalline targets for domains that are amenable to structure determination are underway. We expect to be able to increase our yield of diffraction-quality crystals to at least 50% in the immediate future, because soluble, limit digests of globular proteins represent excellent candidates for crystallization[11–13]. Overcoming this bottleneck will only allow us to find the next rate-limiting step in the pathway depicted in Fig. 3. Our results already suggest that we are facing a problem akin to controlling traffic flow on the island of Manhattan. No sooner is the problem with one intersection resolved, when a traffic jam starts to build at the next choke point, possibly only one block away.

Who should do it and at what cost?

The enormous complexity of the protocol illustrated in Fig. 3 and the collective experience of the various pilot studies suggest that a structural genomics initiative will need to be a centralized endeavour positioned near one or more synchrotron protein crystallography beamlines. While structural genomics pilot studies develop automated tools for cloning, protein expression and purification, crystallization, synchrotron data collection, structure solution/refinement, annotation and dissemination, it makes sense to allow smaller research teams to explore alternate approaches to optimizing the process. Eventually, however, the need to achieve economies of scale and the repetitive nature of the production phase will require integrated centres. In our view, this step is in the best interests of a structural genomics initiative, because it will help address criticisms suggesting that the project is not science and, worse, that it is anti-intellectual. Indeed, many of the criticisms levelled at the Human Genome Project in the mid-1980s have been redirected toward structural genomics.

Perhaps the most bitter aspect of the debate over the Human Genome Project concerned money. Detractors were quick to say that this type of 'big' science would inevitably pauperize investigator-initiated, hypothesis-driven research in molecular and cellular biology. This horrific spectacle did not in fact result from the decision to proceed with high-throughput genome sequencing, and we do not believe that it will happen with structural genomics. NIH officials at the Structural Genomics Workshops (http://www.nih.gov/nigms/funding/psi.html) took great pains to explain that they would not be inclined to fund structural genomics in a way that threatens funding of traditional research grants. Despite considerable discussion, no clear picture has emerged as to how much it will cost to determine 10,000 structures using high-throughput X-ray crystallography. Given the experience of the Human Genome Project, we can predict with certainty that economies of scale will reduce the cost per structure to well below current levels. The NIH-funded pilot studies should provide accurate initial cost estimates, which will be essential for any plausible justification of an expanded effort.

When will the task be substantially complete?

Unlike high-throughput genome sequencing, it is not a simple matter to decide when a structural genomics effort has reached completion. If the analysis uses the uncertain 'new fold' criterion, any definitively stated answer will only serve as a lightening rod for criticism. A more pragmatic approach may be to invoke operational definitions based on the current performance of homology-modelling procedures. After all, one of the most important products of the structural genomics initiative will be homology models. In our pilot study, two protein structures yielded 31 homology models for proteins that were structurally unknown when we made our target selection.

As a general rule, 30% sequence identity represents a critical threshold for successful homology modelling[6,14]. Using this explicit criterion we now consider two fundamentally different situations. When the entire complement of input sequences is known, such as one complete genome or a set of complete genomes, the target list is defined *a priori*. It is a trivial matter to decide when the project is finished. When the input target set is less well defined, such as all genes, it is much more difficult to quantify attainment of completion. If we restrict our analysis to the extreme case of all sequences that we will ever know, the total number of 30% sequence identity protein families is required. Statistical estimates place that number somewhere between 10,000 and 30,000 (C. Sander, unpublished data). The accuracy of this analysis will improve substantially with completion of the Human Genome Project. Thus, it appears that it will be possible to define useful end points for structural genomics, provided one is willing to ask the relevant questions.

What collateral benefits will accrue to biomedicine?

There are four possible outcomes when one determines the structure of a target protein, corresponding to all combinations of "new/old" fold and "known/unknown" function, either its own or that of one or more close homologues. We suggest that the new structural information will be useful in all four cases, but in different ways. At a minimum, every new structure will permit modelling of a protein family for which no structural information is available. Each one of these homology models can serve as a starting point for a rational program of experiments, such as site-directed mutagenesis, ligand-binding studies, enzyme assays, protein-protein interaction studies and so on.

If the structure represents a new fold with a known function, it may well be possible to identify regions of the protein responsible for function *in silico* by comparing the newly determined structure with those of structurally distinct yet functionally similar proteins. This approach has already worked in the context of convergent evolution of enzymes, such as the catalytic triad-bearing serine proteases[15]. When the structure proves to be a known fold with a known function, we can expect to learn something about divergent evolution. This experience has played out repeatedly with the TIM barrel enzymes, which catalyse a wide variety of chemical reactions using the same protein fold decorated with different patterns of surface-accessible residues creating functionally distinct active sites.

Where we do not know anything about biochemical function, both new and previously known structures will still prove useful. The newly determined structures that are not in fact novel can be compared with their structural homologues, and it may be possible to infer function[16]. This was precisely our experience with P007, which most closely resembles an alanine racemase. Regrettably, inferring function from structure does carry some risks. There is no guarantee that the location of the active site in one molecule is of any relevance to the newly characterized protein family. For example, the DNA-binding domains of the transcription factors RFX (K. Gajiwala *et al.*, submitted) and hepatocyte nuclear factor-3γ (ref. 17) share only 8% sequence identity, yet they are both members of the winged-helix family of helix-turn-helix proteins[18]. In the absence of an RFX cocrystal structure, one would have predicted incorrectly that RFX uses its so-called recognition helix to bind DNA. In fact, RFX uses one of its wings to make major groove contacts with its cognate sequence or X-box (K. Gajiwala *et al.*, submitted), making it functionally distinct from all known helix-turn-helix proteins.

The new fold structures for which there is no functional information represent a red flag to some critics of structural genomics. In these cases, however, we have two courses of action open to us. First, novel structures may be functionally characterized by scanning them against a library of all known binding sites and enzyme active sites[15,16]. Second, the new fold structures represent an excellent vantage point from which to develop testable hypotheses regarding function. A useful example can be found in the story of the bacterial periplasmic protein HDEA. Two groups converged on the structure at virtually the same time, but came away with very different views as to its utility (ref. 19, and K. Gajiwala & S.K.B., submitted).

We believe that choosing medically relevant targets will have all of the benefits we have outlined, and in addition a number of important consequences for disease- and patient-oriented research. First, each newly determined structure will be of immediate relevance to academic and/or industrial research teams studying that particular system. Research efforts are invariably accelerated by knowledge of the three-dimensional structures of the biochemical players. Second, by publicizing target lists on the Internet, the structural genomics pilot studies could attract biological and biochemical expertise from the scientific community and even entertain suggestions for additions to their respective target lists. Third, the pilot studies will be able to serve as an important resource for distribution of high-efficiency expression vectors. With sufficient funds, they might produce and distribute large quantities of highly purified proteins for production of antibodies, biochemical and biophysical assays, microinjection into cells, fabrication of arrays of immobilized proteins, high-throughput screening, confirmatory ligand-binding studies and lead-compound optimization. One could imagine that some future NIH grant applications would include both a request for funds and a request for a supply of a particular purified protein deposited in a centralized cold-storage facility. Finally, some of these newly characterized proteins may represent lead compounds in their own right as protein pharmaceuticals.

What will happen to structural biologists if structural genomics becomes a reality? We reject the claim that this endeavour will put X-ray crystallographers and NMR spectroscopists out of business. Instead, we believe that the technological advances that will come from high-throughput structure determinations will improve the efficiency of all structural biologists, making it easier take on the challenges of studying macro-

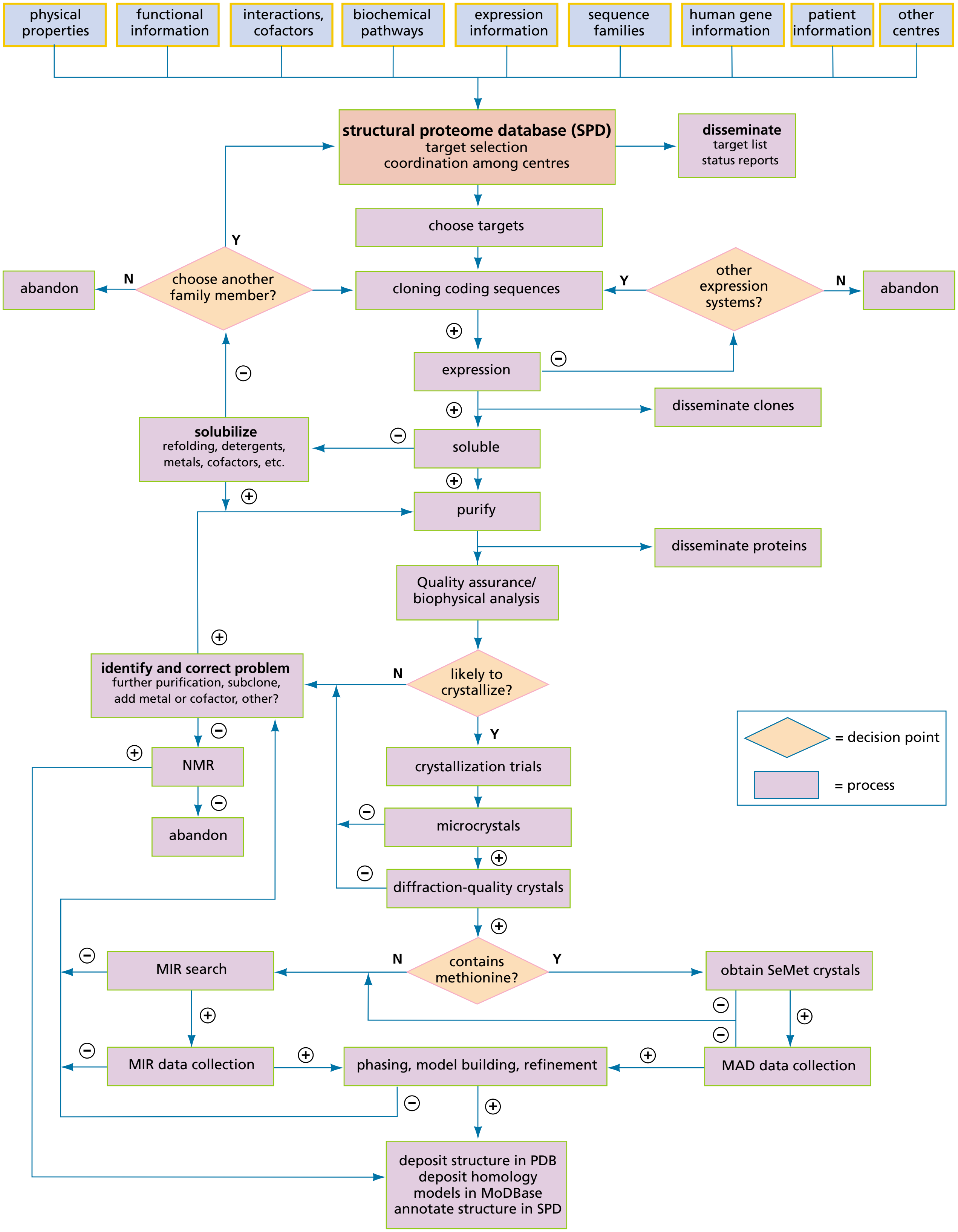

Fig. 3 Flowchart depicting the processes involved in high-throughput structural genomics using X-ray crystallography. (MIR denotes multiple isomorphous replacement, an alternative to MAD for structure determination.)

molecular assemblies (Fig. 2). Moreover, systematic structural studies of all globular domains will in no way exhaust protein sequence space. The problem of membrane protein crystallization remains vexing, rendering them very unlikely candidates for a structural genomics target list. Not to mention the problem of trying to understand the function of proteins that remain unstructured until they interact with their targets.

By way of conclusion, it is worth noting that structural genomics has the potential to go beyond the practical benefits outlined above. One of the great unsolved problems in molecular biology concerns the relationship between one dimensional sequence information and three-dimensional structure—the protein folding problem. Anfinsen won the Nobel Prize in Chemistry for demonstrating that all of the information required to determine the fold of a protein is contained in the order of amino acids comprising the polypeptide chain. Some workers have referred to this phenomenon as the next genetic code, but it has proved much harder to crack than the triplet code underpinning protein translation. Structural genomics may well provide the means of coming to grips with this important intellectual challenge.

Acknowledgements

This work was supported in part by the National Institutes of Health Center for Research Resources, the Department of Energy, Albert Einstein College of Medicine and The Rockefeller University. S.K.B. is an Investigator in the Howard Hughes Medical Institute.

1. Brenner, S.E., Chothia, C. & Hubbard, T. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376 (1997).
2. Orengo, C.A. *et al.* The CATH Database provides insights into protein structure/function relationship. *Nucleic Acids Res.* **27**, 275–279 (1999).
3. Fischer, D. & Eisenberg, D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. *Proc. Natl Acad. Sci. USA* **94**, 11929–11934 (1997).
4. Rychlewski, L., Zhang, B. & Godzik, A. Fold and function predictions for Mycoplasma genitalium proteins. *Fold. Des.* **3**, 229–238 (1998).
5. Huynen, M. *et al.* Homology-based fold predictions for Mycoplasma genitalium proteins. *J. Mol. Biol.* **280**, 323–326 (1998).
6. Sanchez, R. & Sali, A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA* **95**, 13597–13602 (1998).
7. Gaasterland, T. Structural genomics taking shape. *Trends Genet.* **14**, 135 (1998).
8. Sali, A. 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032 (1998).
9. Studier, F.W., Rosenberg, A.H., Dunn, J.J. & Dubendorff, J.W. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185**, 60–89 (1990).
10. Holm, L. & Sander, C. A review of the use of protein structure comparison in protein classification and function identification. *Science* **273**, 595–602 (1996).
11. Cohen, S.L., Ferre-D'Amare, A.R., Burley, S.K. & Chait, B.T. Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. *Protein Sci.* **4**, 1088–1099 (1995).
12. Cohen, S.L. Domain elucidation by mass spectrometry. *Structure* **4**, 1013–1016 (1996).
13. Xie, X. *et al.* Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature* **380**, 316–322 (1996).
14. Sanchez, R. & Sali, A. Comparative proteins structure modeling in genomics. *J. Comp. Phys.* **151**, 388–401 (1999).
15. Wallace, A.C., Borkakoti, N. & Thornton, J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323 (1997).
16. Fetrow, J.S., Godzik, A. & Skolnick, J. Functional analysis of the Eschericia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711 (1998).
17. Clark, K.L., Halay, E.D., Lai, E. & Burley, S.K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412–420 (1993).
18. Lai, E., Clark, K.L., Burley, S. & Darnell, J.E. Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: a family of transcription factors of diverse biological function. *Proc. Natl Acad. Sci. USA* **90**, 10421–10423 (1993).
19. Yang, F., Gustafson, K.R., Boyd, M.R. & Wlodawer, A. Crystal structure of Escherichia coli Hdea. *Nature Struct. Biol.* **5**, 763–764 (1998).
20. Guenther, B.D., *et al.* The structure and properties of methylenetetrahydrofolate reductase from Escherichia coli suggest how folate ameliorates human hyperhomocysteinemia. *Nature Struct. Biol.* **6**, 359–365 (1999).
21. Deo, R.C., Bonanno, J.B., Sonenberg, N. & Burley S.K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* (in press).